

# Opinion Mining and Sentiment Analysis Need Text Understanding

Rodolfo Delmonte and Vincenzo Pallotta

**Abstract.** We argue in this paper that in order to properly capture opinion and sentiment expressed in texts or dialogs any system needs a deep linguistic processing approach. As in other systems, we used ontology matching and concept search, based on standard lexical resources, but a natural language understanding system is still required to spot fundamental and pervasive linguistic phenomena. We implemented these additions to VENSES system and the results of the evaluation are compared to those reported in the state-of-the-art systems in sentiment analysis and opinion mining. We also provide a critical review of the current benchmark datasets as we realized that very often sentiment and opinion is not properly modeled.

## 1 Introduction

Sentiment analysis and Opinion mining [Pang and Lee 2008] are emerging applications of Natural Language Processing whose importance is becoming increasingly higher. Brand managers and market researchers use these applications in order to monitor the “voice of the customer” in order to study trends of acceptance/rejections and sometimes to discover issues in products or services.

We assume that in order to properly capture opinion and sentiment expressed in a text or dialog any system needs a full natural language understanding (NLU) approach. In particular, the idea that the task may be solved by the use of

---

Rodolfo Delmonte  
Department of Language Science  
Università “Ca Foscari”  
Dorsoduro, 3462 - Venezia  
30123 – Venezia, Italy  
e-mail: delmont@unive.it

Vincenzo Pallotta  
Department of Informatics  
University of Fribourg  
Bd. des Pérolles, 90,  
1700 - Fribourg, Switzerland  
e-mail: vincenzo.pallotta@unifr.ch

Information Retrieval tools like Bag of Words (BOWs) approaches has shown its intrinsic shortcomings. In fact, in order to achieve acceptable results, BOWs approaches are sometimes camouflaged by a keyword-based Ontology matching and Concept search, based on SentiWordNet<sup>1</sup> [Bentivogli et al. 2004], by simply stemming a text and using content words to match its entries and produce some result. Any search based on keywords and BOWs is fatally flawed by the impossibility to cope with such fundamental and pervasive linguistic phenomena as the following ones:

- presence of NEGATION at different levels of syntactic constituency;
- presence of LEXICALIZED NEGATION in the verb or in adverbs;
- presence of conditional, counterfactual subordinators;
- double negations with copulative verbs;
- presence of modals and other modality operators.

In order to cope with these linguistic elements we propose to build a Flat Logical Form (FLF) directly from a Dependency Structure representation of the content augmented by indices and where anaphora resolution has operated pronoun-antecedent substitutions. We implemented these additions our NLU system called VENSES [Delmonte et al. 2009]. The output of the system is an XML representation where each sentence of a text or dialog is associated to a list of attribute-value pairs, one of which is POLARITY. In order to produce this output, the system makes use of the FLF and a vector of semantic attributes associated to the verb at propositional level and memorized.

Important notions such as the distinction of the semantic content of each proposition into two separate categories, OBJECTIVE vs. SUBJECTIVE, are also required by the computation of opinion and sentiment. This distinction is obtained by searching for FACTIVITY markers again at propositional level. In particular we take into account:

- tense, voice, mood at verb level
- modality operators like intensifiers and diminishers, but also modal verbs
- modifiers and attributes adjuncts at sentence level
- lexical type of the verb (in Levin's classes and also using WordNet classification)
- subject's person (if 3<sup>rd</sup> or not).

The article is organized as follows. In section 2, we review the components of the VENSES system. In section 3 we present its tailoring for the task of sentiment analysis. Section 4 describes the experiment we carried out on a benchmark dataset where we compare our results to the state-of-the-art results and discuss the flaws of BOW systems. Section 5 concludes the paper with some lesson learned and recommendations.

---

<sup>1</sup> <http://sentiwordnet.isti.cnr.it/>

## 2 The VENSES System

VENSES is a tailored version of GETARUNS<sup>2</sup> [Delmonte 2007; 2008b], a complete system for text understanding developed at the Laboratory of Computational Linguistics of the University of Venice. The system produces different levels of analysis, from syntax to discourse. However, three of them contribute most to the success of sentiment analysis:

1. the syntactic and lexico-semantic module,
2. the anaphora resolution module [Delmonte et al. 2007],
3. the deep semantic module.

### 2.1 The Syntactic and Lexico-Semantic Module

GETARUNS, is organized as a pipeline which includes two versions of the system: what we call the *Deep* and *Partial* GETARUNS.

The Deep version of GETARUNS is equipped with three main modules: a *lower module* for parsing, where sentence strategies are implemented; a *middle module* for semantic interpretation and discourse model construction which is cast into Situation Semantics; and an *upper module* where reasoning and generation takes place.

GETARUNS, has a highly sophisticated linguistically based semantic module which is used to build up the DM. Semantic processing is strongly modularized and distributed amongst a number of different sub-modules, which take care of Spatio-Temporal Reasoning, Discourse Level Anaphora Resolution, and other subsidiary processes like Topic Hierarchy.

The architecture of the Partial GETARUNS is shown in Figure 1. This version is fired before the Deep system and is used as a back-off strategy whenever failures ensue. The Partial system tries at first to produce a full parse of the current utterance with the lower system. The Deep system makes use of chunks as produced by the Partial system, in order to guess where in the utterance is positioned the current analysis, in particular where the VP starts. Only in case the Deep system fails, the Partial system will proceed by producing Partial semantics and Discourse level analysis through middle and upper level.

The parser produces a c-structure representation by means of a cascade of augmented FSA<sup>3</sup>. Then it uses this output to map lexical information from a number of different lexica, which however contain similar information related to verb/adjective and noun sub-categorization. The mapping is done by splitting the

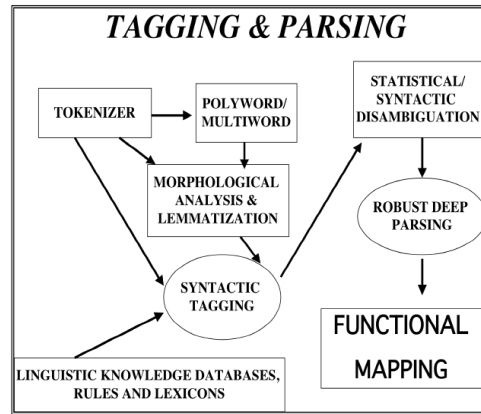
<sup>2</sup> The system has been tested in STEP competition (see [Delmonte 2008a], and can be downloaded in two separate places. The partial system called VENSES in its stand-alone version is available at:

[http://www.aclweb.org/aclwiki/index.php?title=Textual\\_Entailment\\_Resource\\_Pool](http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool)

The complete deep system is available both at:

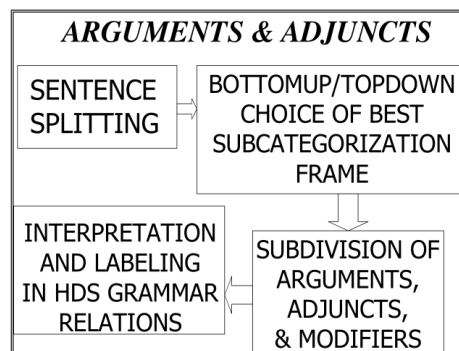
<http://project.cgm.unive.it/html/sharedtask/>

<sup>3</sup> Finite State Automata.



**Fig. 1** GETARUNS: lower level.

sentences into clauses, which are normally main and subordinate clauses. Other clauses are computed in their embedded position and can be either complement or relative clauses.



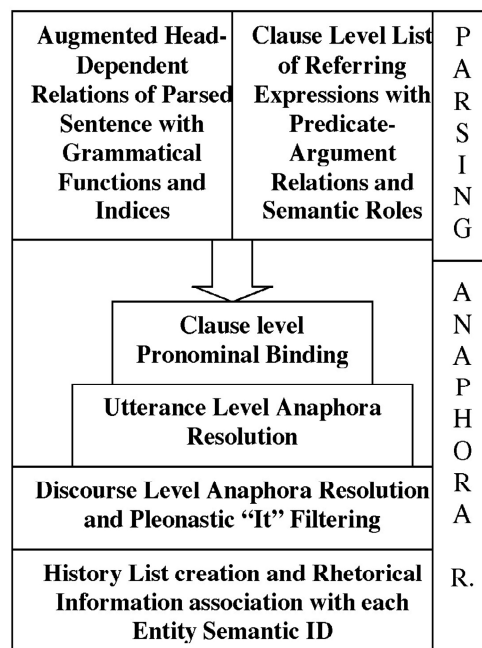
**Fig. 2** GETARUNS: upper level.

The output of the upper level is what we call AHDS (Augmented Head Dependent Structure), which is a fully indexed logical form, with Grammatical Relations and Semantic Roles. The inventory of semantic roles we use is however very small (i.e. 35) even though it is partly overlapping with the set proposed in the first FrameNet project<sup>4</sup>. We prefer to use Generic Roles rather than specific Frame Elements (FEs) because sense disambiguation at this stage of computation may not be effective.

<sup>4</sup> <http://framenet.icsi.berkeley.edu/book/book.pdf>

## 2.2 The Anaphora Resolution Module

This module whose components are sketched in Figure 3 works on the so-called History List of entities present in the text so far. In order to make the output of this module usable by the Semantic Evaluator, we decided to produce a flat list of semantic vectors which contain all semantic related items of the current sentence. Inside these vectors, pronominal expressions are substituted by the heads of their antecedents.



**Fig. 3** GETARUNS: the anaphora module

The AHDS structure is passed to and used by a full-fledged module for pronominal and anaphora resolution, which is in turn split into *two sub-modules*. The resolution procedure takes care only of third person pronouns of all kinds – reciprocals, reflexives, possessive and personal. Its mechanisms are quite complex and details can be found in [Delmonte et al. 2006]. The *first sub-module* basically treats all pronouns at sentence level – that is, taking into account their position – and if they are left free, they receive the annotation “external”. If they are bound, they are associated to an antecedent’s index; else they might also be interpreted as expletives, i.e. they receive a label that prevents the following sub-module to consider them for further computation.

The *second sub-module* receives as input the external pronouns, and tries to find an antecedent in the previous stretch of text or discourse. To do that, the system computes a *topic hierarchy* that is built following suggestions by [Grosz and Sidner 1986] and is used in a centering-like manner.

### 2.3 The Deep Semantic Module

The output of the anaphora resolution module is fed to the deep semantic module in order to substitute the pronoun's heads with the antecedent's heads. After this operation, the semantic module produces Predicate-Argument Structures (PASs) on the basis of previously produced Logical Form. PASs are produced for each clause and they separate mandatory from non-mandatory arguments, and these from adjuncts and modifiers. Some adjuncts, like spatio-temporal locations, are only bound at propositional level.

This module also produces a representation at propositional level, which for simplicity is just a simple vector of information containing 15 different slots, each one containing a different piece of relevant semantic information. We encode the following items: modality, negation, focusing intensifiers/diminishers, manner adjuncts, diathesis, auxiliaries, clause dependency if any from a higher governing predicate – this is the case for infinitivals and gerundives – and a subordinator, if any.

## 3 The Sentiment Analysis System

Differently from other sentiment analysis systems, we use a three-way classification for the attribute “attitude” which encodes polarity: POSITIVE, NEGATIVE and SUSPENSION. The SUSPENSION category is used when negation is present in the utterance but the overall attitude is not directly negative.

[Hu and Liu 2004] uses a scale of three grades to indicate strength and distinguish cases of real NEGATIVE/POSITIVE polarity from one another. In the readme file associated to the datasets, the authors comment on this grading system that: “... *note that the strength is quite subjective. You may want to ignore it, but only considering + and -*”.

It is a fact that annotation criteria are hard to establish, but then the outcome will always be subjective in a sense. For instance in the example below taken from one of the datasets made available by the authors and on which we evaluated our system, the score [-1] indicates low negative polarity strength.

In order to have an idea of where the problems lie, we report below how the example was annotated in the dataset (A), and then the output of our system (B):

```
A. viewfinder[-1]##the lens is visible in the viewfinder when the lens is set to the wide
angle , but since i use the lcd most of the time , this is not really much of a bother to me.
B. id="44", predicate="be", topic="lens", attitude="suspension" factivity= "factive_
statement"
```

It is apparent that this example cannot be considered as fully negative and probably the negative category alone does not capture the nuances of the statement. We take a conservative approach and we classify this statement as being SUSPENSION because negative expressions cannot be directly mapped onto a negative opinion. Here and elsewhere we annotated SUSPENSION, and the system correctly labels the example: this label indicates an attitude which is not strongly marked for either polarity value, and in some cases this may also be due to the presence of double negation. As a consequence, we also use SUSPENSION in the following example, where (Hu and Liu 2004) annotates instead as Positive with high confidence:

A. weight[+2]##at 8 ounces it is pretty light but not as light as the ipod .  
 B. id="46", predicate="be", topic="light", attitude="suspension" factivity= "factive\_statement"

It is apparent that this case cannot be computed as a strong case for positive attitude, actually this is not a positive opinion at all. We would like to stress here that those shown in A are manual annotations and not output from an automatic classification system. This also entails that the training dataset is intrinsically flawed due to a misunderstanding of what polarity actually is.

In many other cases, Hu and Liu's provide no annotation, which does not mean, in our opinion, that the utterance can be considered as neutral, as for instance in the following cases:

A. ##if you have any doubts about this player , well do n't .  
 B. id="32", predicate="have" topic="player" attitude="suspension" factivity= "opinion\_internal"

A. ##can 't complain and i recommend it over all the other players , just hope that remote will come out soon .  
 B. id="27", predicate="recommend", topic="player", attitude="suspension", factivity= "opinion/ factive\_statement"

In many cases, however, it is hard to understand the reason why the annotation has not been made available for trivial cases as for instance in,

- "do not buy this player"
- "a piece of junk"
- "don't waste your money".

### 3.1 *Sentiment Analysis of Conversations*

We tailored our system to deal with contexts larger than short reviews of products. Here below we present an excerpt from a short dialog, which contains a certain number of complex negative cases to solve. In Table 1 we provide our analysis of each sentence:

**Table 1** Sentiment and Factivity analysis of a conversation excerpt.

id	Content	Predicate	Polarity	Factivity
1	Well, what do you think?	say	positive	question
2	That's not so bad.	be	suspension	opinion_internal
3	I'm not complaining.	complain	suspension	opinion_statement
4	That's not true.	be	negative	opinion_statement
5	Jack never contradicts my opinions.	contradict	suspension	factive_statement
6	Mark always contradicted my ideas.	contradict	negative	factive_statement
7	Mark never accepted disadvantages.	accept	negative	factive_statement
8	Nobody bought that product.	buy	negative	factive_statement
9	I bought an awful product.	buy	negative	factive_statement
10	I don't like that product.	like	negative	factive_statement
11	I strongly criticize such a product.	criticize	negative	factive_statement
12	No sensible customer would buy that product.	buy	negative	factive_statement
13	Mary bought that product for an awful purpose.	buy	negative	factive_statement
14	Mary bought that product to kill herself.	buy	negative	factive_statement
15	Mark didn't make a bad deal.	make	relevant	opinion_statement
16	That product doesn't seem to be awful.	seem	suspension	opinion_internal
17	Mary didn't buy that awful product.	buy	negative	factive_statement
18	John didn't kill the bad feelings of the customers about that awful product.	kill	suspension	factive_statement

As it can be easily noticed, the problem is due to the presence of negation that is not solvable by a simple one-way decision – yes/no. In many cases the information about the attitude of the speaker is just not directly communicated and needs further specification. In sentence 16, for instance, is not a straightforward admission of disagreement; the same applies to sentence 18. We also regard sentences 2, 3, 5 to be cases of *indirect judgment*, which however is not explicit enough to be assigned to a positive attitude. For this reason, we decided to introduce the SUSPENSION marker, which encodes all cases of indirect judgment, and other similar situations.

Coming now to clear cases of NEGATIVE attitude, we register sentences like 4, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 17. However, not all these sentences can be easily understood as being totally NEGATIVE. In particular, only sentence 4 and 10 are simple cases of negation at main verb level and may be computed safely as cases of negative attitude. Sentences 6 and 11 are again cases of negative attitude but there is no explicit negation expressed: just negatively marked verb at lexical level.

Examples 8 and 12 express negation at subject level: as can be gathered, this can only be evaluated as a real negative attitude only in case the main verb indicates positive actions. Apparently, these cases can also be contradicted by the same speaker, by using BUT and other adversative discourse markers (e.g. “even though nobody likes it...”; “nobody likes it”, “but ...”). Examples 9, 13, and 14



introduce negative attributes at object and complement level. This is also computed by the system as a case of negative attitude.

The system also computes as negative example 17, which is a case of double negation: in this sentence, negation is present both at verbal level and at complement level. This might be understood as positive attitude (i.e. “if she did not do that then it is good...”). However we assume that this is also interpretable as a report of something negative that might have happened and not as a negative judgment. This distinction may seem subtle, but we believe is very important in order to avoid false positives in the classification. Eventually, we have also cases of SUSPENSION involving the presence of negation as example 18 shows.

An important subdivision of all semantic types involved, regards FACTIVITY, which, as we said before, can constitute an important indicator of the speaker’s attitude in uttering a given judgment. It is important to mention that sentences are in fact utterances that can be categorized in at least two main types:

1. they constitute OBJECTIVE (or FACTIVE) STATEMENT reporting in this way some fact usually in third person subject;
2. they may constitute SUBJECTIVE (or NON-FACTIVE) OPINIONS expressed by the speaker him/herself in first person or reporting on somebody else’s opinion.

Opinions are always subjective but may also report an internal thought, a wish, a hope, or else a definite state, event, and activity by the subject. In the former case, we use OPINION\_INTERNAL, to highlight the weight of subjective markers as in 2 and 16. In the latter case, we use OPINION\_STATEMENT because it is either the case that the utterance refers acts or events of third persons, as in 15; or else, it reports the evaluation of the speaker as in 3 and 4. Other markers are QUESTION, which can still be computed as either positive or negative; and RELEVANT, implying some indirect judgment as shown by 16 and double negation and reinforcing on SUSPENSION.

As a last example, we now consider a really difficult utterance to evaluate from Hu and Liu’s dataset:

\*Positive-1 dvd - so far the dvd works so i hope it doesn't break down like the reviews i 've read.

This item has been correctly annotated as POSITIVE in Hu and Liu’s dataset. However, in order to capture the dependency between the negated sentence and the “hope” predicate, a system definitely needs to build a logical form and all the appropriate indices. Predicates like “hope” make the following governed proposition “opaque” and non-factive. This forces the system to “dummify” the presence of negation. Looking carefully at the example, this sentence cannot be considered as positive as it is a clear case of SUSPENSION where no judgment has been expressed, but only a worry that other reviews made would turn out to be true in reality.

### 3.2 *The Semantic Markers: CONDITIONAL and COMPARATIVES*

Eventually, there are important components of a semantic analysis, which may heavily influence the final output. We are now referring to two well-known cases discussed in the literature: the presence of “conditional” discourse markers like IF, WHETHER which transform a statement into a conditional clause which is usually accompanied by the presence of “unreal” mood like conditional or subjunctive. And then we come to “comparative” constructions, which are more frequent in consumer product reviews than in blogs or social networks opinions. As far as comparatives are concerned, it is a fact that real utterances contain a gamut of usage of such a construction, which is very hard to come to terms with. We list some of the most relevant cases here below and then make some comments. Each utterance is taken from Hu and Liu’s reviews databases and has an evaluation at the beginning of the line:

- a. \*Positive-2 player - i did not want to have high expectations for this apex player because of the price but it is definitely working out much better than what i would expect from an expensive high-end player.
- b. \*Positive-2 look - without a doubt the finest looking apex dvd player that i 've seen.
- c. \*Positive-2 dvd player - so sit back , relax and brag to all your friends who paid a mountain of money for a dvd player that can't do half the things this one can , and for a fraction of the price !
- d. \*Positive-3 camera - recent price drops have made the g3 the best bargain in digital cameras currently available.
- e. \*Positive-2 feel - you feel like you are holding something of substance , not some cheap plastic toy.
- f. \*Positive-3 camera - i can't write enough positive things about this great little camera !
- g. \*Positive-3 camera - this is my first digital camera and i couldn't be happier.
- h. \*Positive-3 finish - its silver magnesium finish is stunning, and the sharp lines and excellent grip are better than any other camera i've seen.
- i. \*Positive-2 noise another good thing is that this camera seems to introduce much less noise in dark places than others i've seen.
- k. \*Positive-2 camera this is by far the finest camera in its price and category i have ever used.

As it can be noticed, in many cases what is really the guiding principle is the need of comparing the evaluative content of two opposing propositions, rather than simply measuring degree of comparison (i.e. superlative rather than comparative grade). In example a. the first proposition is negated and then the second compared proposition marked by BUT is a really hard complex sentence to compute. In b. one has to compute correctly “without a doubt”. In c. the first proposition has a relative clause referring to a negative fact, where however the governing verb BRAG can be understood both negatively and positively. In d. the phrase “recent price drops” can be a negative fact but has to be understood positively together with the following proposition where “best bargain” appears. Again in e. one needs to compare two propositions one of which has an ellipsed VP. In f. the

reviewer uses a rhetorical device “can’t write enough positive...” which however introduces negation. The same applies to example g.

#### 4 The Experiment with Products Reviews

In order to evaluate our system, we used [Hu and Liu 2004] datasets, which have been collected and partially annotated in 2004 and are made of customer reviews of 5 products downloaded from Amazon.com. In fact, we used for perusal and evaluation only three of them: Canon (digital camera), Creative (mp3 player) and Apex (dvd player). The problem was that the annotated examples were just a small percentage of the total - 1302 sentences over 3300, so we had to manually annotate the remaining cases (60% of all utterances) ourselves and make some corrections on the input: the texts were full of typos and had many non-words, fragments, ungrammatical sentences etc. Overall, we parsed 30,000 tokens and 3300 utterances. In Table 2 we report some statistics about the three datasets we have used in our experiment. The Sents column indicates the number of total utterances present in each dataset.

**Table 2** Annotation data from Ho and Liu’s datasets

	Positive	Negative	Totals	Sents
apex	148	195	343	840
canon	184	54	238	643
creative	421	299	720	1811
totals	753	548	1301	3394

As can be easily noticed, only 38.33% (1301 out of 3394) of all utterances have been annotated, which makes the comparison fairly difficult to make. In particular, if we look at our annotated data in Table 3, the overall number of NEGATIVE polarity judgments constitutes 58% of all judgments when compared to 42% in Ho and Liu’s annotations. The final outcome is then totally mistaken: in our case the judgments are more negative and in Ho and Liu’s they are more positive disregarding the subdivision of reviews into each separate products. We computed the number of annotations in original datasets, which have been graded [+/- 1], thus indicating that the confidence of the annotator is very low, and this makes up 16.37% of all annotations. In our case, the SUSPENSION annotations constitute 23.22% of all annotations.

**Table 3** Automatic annotation with VENSES

	Pos.	Neg.	Susp.	Quest.	Totals
apex	327	300	199	15	841
canon	294	197	143	13	647
creative	558	782	430	37	1797
totals	1030	1447	769	65	3311

The first interesting fact to notice is the slight difference in Recall, where we see that of all the utterances present we only got 97.55%. It is important to highlight the difference in the approach. Our system's output refers to *real utterances*, which sometimes do not coincide with each line or record in the input file. The system computes an utterance every time it finds a sentence delimiting punctuation mark. As a result, in some cases, as in "canon" dataset, we end up with additional utterances to evaluate.

**Table 4** Evaluation on the basis of Hu and Liu's gold standard

	Accuracy	Accuracy %	F-score
apex	286/343	83.39	90.94%
canon	174/238	73.00	84.39%
creative	547/720	76.20	86.49%
totals	1007/1301	77.40	87.26%

The results of the evaluation shown in Table 4 are based on Hu and Liu's dataset at first and are computed for accuracy as a ratio of correct/gold standard; we also compute the F-score<sup>5</sup>, where Recall is in our case equal to 100% in the sense that we compute all evaluations for all sentences.

**Table 5** Evaluation on the basis of our annotation

	Precision	Recall	F-score
apex	670/841=79.66%	826=98.41%	88.05%
canon	468/648=72.23%	638=98.45%	83.32%
creative	1424/1811=78.63%	1764=97.40%	87.01%
totals	2562/3300=77.60%	3228/3300=97.81%	86.54%

Table 5 shows results computed on the basis of the overall annotation integrated by our single annotator, has a slightly lower F-score. It is interesting to note the difference in overall polarity evaluation, which may affect the opinion of prospective customers inducing them in buying or not buying a certain product on the basis of the balance between positive and negative polarity. In the evaluation carried out by Ho and Liu we see that negative judgments constitute the 42.12% (548 negative and 753 positive). In our case the proportions are reverse: we have (1030 positive and 1447 negative) 58.42% negative judgments.

Data related to SUBJECTIVITY and FACTIVITY reported in Table 6 show a balanced subdivision of all data between the two categories.

<sup>5</sup> The F-score is  $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$ .

**Table 6.** Subjectivity results from VENSES

	Fact	Opin	Opin_ Inter	Fact/Opin	Total
apex	398	265	87	100	850
canon	300	226	47	75	648
creative	772	609	195	219	1795
totals	1470	1100	329	394	3293

#### 4.1 *The Experiment with Quotations*

We did another experiment on the basis of a corpus of news annotated at the JRC of the European Commission [Balahur et al. 2010]. The corpus is a collection of 1592 quotations, which however have been collected automatically. It comes out that they contain 151 totally repeated texts, 13 non-sentences or even non-fragments that cannot be evaluated at all. Then there are some 24 quotes which are constituted by questions, which again not being statements cannot be evaluated negatively/positively. Another 15 quotations are portions of quotes and have been included in the evaluation. At the end we came up with 1404 quotations. In fact, as the authors report in their paper, only 1292 are fully agreed quotes on the basis of their three annotators. The evaluation the authors present at the end of their paper is based on a small amount of data - 427 quotes - constituted only by those quotes, which have received full agreement in their polarity judgments. The authors leave out 865 quotes, which have been computed as objective statements, on the assumption that only subjective quotes can be regarded appropriate for polarity judgment.

We don't agree at all with their definition of polarity judgment: "statements may describe objective negative state of affairs, current or even future events much in the same way in which subjective statements do". In fact, since quotations are mainly third person descriptions, narrations, or reported speech they can belong to both categories.

The results in terms of accuracy are similar to those on reviews and dialogs: 82.05% for negative judgments, 70.34% for positive judgments, overall 76.2% accuracy. The main difference is constituted by the evaluation of positive judgments, which are indirectly reported and require a lot of semantic knowledge. Consider quotes like the following ones that are evaluated for positive:

1. "anybody who wants [Mr Obama] to fail is an idiot, because it means we're all in trouble..."
2. "Charles Freeman was the wrong guy for this position. His statements against Israel were way over the top and severely out of step with the administration. I repeatedly urged the White House to reject him, and I am glad they did the right thing."

The first sentence is correctly classified as non-negative with respect to the main topic, Obama, as well as for the second, which is a reply to a negative comment and it should thus count as positive. These cases are clearly hard to capture for a

system without a deep language understanding and that is capable to deal with larger context than a single unit.

## 5 Conclusions

In this paper we have advocated the need of a Natural Language Understanding system to adequately deal with the task of sentiment analysis and opinion mining. We pointed out the issues in both annotated datasets used for benchmarking and the mainstream methods that are based on enhanced Bag of Word approaches. Our conclusions that Sentiment Analysis and Opinion Mining community needs better datasets and more precise annotation guidelines.

For what concern the performance of our system, we found that not being based on Machine-Learning, it can be substantially improved on the basis of better rules and better lexica. In fact, most mistakes are due to the presence of wrong polarity assignments in the lexical resources used such as the Harvard's General Inquirer dictionary<sup>6</sup> or the Wiebe's list [Wiebe and Mihalcea 2006]. In fact, what can apply to psychology tests does not always apply to the evaluation of reviews, which have products as their objects. In addition, we discovered that there are a variable number of cue words that are ambiguous and vary their connotation (i.e. from positive may become negative and vice versa) according the domain of application. In the field of photography for instance, words such as shoot or shot do not carry negative connotation. So eventually, the system must be updated with respect to the domain and this is something that can be done using WordNet Domains, a resource made freely available by IRST/FBK (see [Bentivogli et al. 2004]), which indicates domains with the needed perspicuity.

## References

- [Balahur et al. 2010] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, May 19-21, pp. 2216–2220 (2010)
- [Bentivogli et al. 2004] Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing. In: Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, pp. 101–108 (2004)
- [Delmonte 2007] Delmonte, R.: Computational Linguistic Text Processing – Logical Form, Logical Form, Semantic Interpretation, Discourse Relations and Question Answering. Nova Science Publishers, New York (2007)
- [Delmonte et al. 2007] Delmonte, R., et al.: Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach. In: Proceedings of ROMAND 2006 - 11th EACL, Geneva, pp. 3–10 (2006)

---

<sup>6</sup> <http://www.wjh.harvard.edu/~inquirer/homecat.htm>

- [Delmonte 2008a] Delmonte, R.: Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution. Nova Science Publishers, New York (2008)
- [Delmonte 2008b] Delmonte, R.: Semantic and Pragmatic Computing with GETARUNS. In: Bos, Delmonte (eds.) Proceedings of Semantics in Text Processing (STEP), Research in Computational Semantics, vol. 1, pp. 287–298. College Publications, London (2008)
- [Delmonte et al. 2009] Delmonte, R., Tonelli, S., Tripodi, R.: Semantic Processing for Text Entailment with VENSES. In: Proceedings of the TAC 2009 Workshop on TE, Gaithersburg, Maryland (November 17, 2009)
- [Hu and Liu 2004] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of KDD 2004 (2004)
- [Pang and Lee 2008] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2) (2008)
- [Grosz and Sidner 1986] Grosz, B., Sidner, C.: Attention, intentions, and the structure of discourse. Computational Linguistics 12(3), 175–204 (1986)
- [Wiebe and Mihalcea 2006] Wiebe, J., Mihalcea, R.: Word Sense and Subjectivity. In: Proceedings of ACL 2006 (2006)

